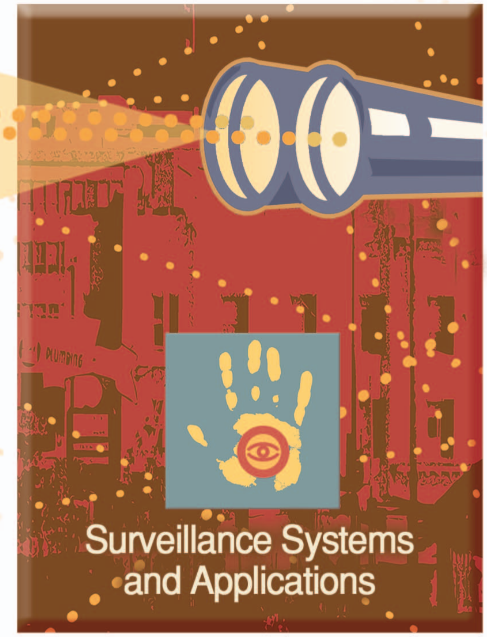
Arun Hampapur, Lisa Brown, Jonathan Connell, Ahmet Ekin, Norman Haas,
Max Lu, Hans Merkl, Sharath Pankanti, Andrew Senior, Chiao-Fe Shu, and Ying Li Tian



© EYEWIRE

# Smart Video Surveillance

## Exploring the concept of multiscale spatiotemporal tracking

Situation awareness is the key to security. Awareness requires information that spans multiple scales of space and time. A security analyst needs to keep track of "who are the people and vehicles in a space?" (identity tracking), "where are the people in a space?" (location tracking), and "what are the people/vehicles/objects in a space doing?" (activity tracking). The analyst also needs to use historical context to interpret this data. For example, the fact that the paper delivery truck showed up at 6 a.m. instead of the usual 8 a.m. would alert a security analyst.

Smart video surveillance systems are capable of enhancing situational awareness across multiple scales of space and time. However, at the present time, the component technologies are evolving in isolation; for example, face recognition technology addresses the identity tracking challenge while constraining the subject to be in front of the camera, and intelligent video surveillance technologies provide activity detection capabilities on video streams while ignoring the identity tracking challenge. To provide comprehensive, nonintrusive situation awareness, it is imperative to address the challenge of multiscale, spatiotemporal tracking. This article explores the concepts of multiscale spatiotemporal tracking through the use of real-time video analysis, active cameras, multiple object models, and long-term pattern analysis to provide comprehensive situation awareness.

### INTRODUCTION

Ensuring high levels of security at public access facilities like airports and seaports is an extremely complex challenge. A number of technologies can be applied to various aspects of the security challenge, including

1053-5888/05/$20.00©2005IEEE

screening systems for people and objects (bags, vehicles, etc.), database systems for tracking "trusted people," biometric systems to verify identity, and video surveillance systems to monitor activity. Today's video surveillance systems act as large-scale video recorders, either analog or digital. Their primary focus is the application of video compression technology to efficiently multiplex and store images from a large number of cameras onto mass storage devices (either video tapes or disks). These systems serve two key purposes: providing a human operator with images to detect and react to potential threats and recording evidence for investigative purposes. While these are the first steps in using video surveillance to enhance security, they are inadequate for supporting both real-time threat detection and forensic investigation.

From the perspective of real-time threat detection, it is a well-known fact that human visual attention drops below acceptable levels even when trained personnel are assigned to the task of visual monitoring [4]. From the perspective of forensic investigation, the challenge of sifting through large collections of surveillance video tapes is even more tedious and error prone for a human investigator. Automatic video analysis technologies can be applied to develop smart surveillance systems that can aid the human operator in both real-time threat detection and forensic investigatory tasks. Specifically, multiscale tracking technologies are the next step in applying automatic video analysis to surveillance systems.

In this article, we begin with a discussion on the state-of-the-art in video analysis technologies as applied to surveillance and the key technical challenges. Component technologies for a smart surveillance system are then presented. We explore a face cataloging system and discuss a long-term site monitoring and movement pattern analysis system. The article concludes with a discussion of future directions.

## STATE-OF-THE-ART IN VIDEO ANALYSIS FOR SURVEILLANCE

Video analysis and video surveillance are active areas of research. The key areas are video-based detection and tracking, video-based person identification, and large-scale surveillance systems. A significant percentage of basic technologies for video-based detection and tracking were developed under a U.S. government-funded program called Video Surveillance and Monitoring (VSAM)[2]. This program looked at several fundamental issues in detection, tracking, autocalibration, and multicamera systems [8], [9], [19]. There has also been research on real-world surveillance systems in several leading universities and research labs [16]. The next generation of research in surveillance is addressing not only issues in detection and tracking but also issues of event detection and automatic system calibration [20].

The second key challenge of surveillance—namely, video-based person identification—has also been a subject of intense research. Face recognition has been a leading modality with both ongoing research and industrial systems [1], [13]. A recent U.S. government research program called Human ID at a Distance addressed the challenge of identifying humans at a distance using techniques like face at a distance and gait-based recognition [10].

One of the most advanced systems research efforts in large-scale surveillance systems is the ongoing U.S. government program titled Combat Zones That See [3]. This program explores rapidly deployable smart camera tracking systems that communicate over ad hoc wireless networks, transmitting track information to a central station for the purposes of activity monitoring and long-term movement pattern analysis.

## *KEY CHALLENGES*

There are several technical challenges that need to be addressed to enable the widespread deployment of smart surveillance systems. Of these, we highlight three key challenges.

### THE MULTISCALE CHALLENGE

One of the key requirements for effective situation awareness is the acquisition of information at multiple scales. A security analyst who is monitoring a lobby observes not only where people are in the space and what they are doing but also pays attention to the expression on people's faces. The analyst uses these visual observations in conjunction with other knowledge to make an assessment of threats. While existing research has addressed several issues in the analysis of surveillance video, very little work has been done in the area of better information acquisition based on real-time automatic video analysis, like automatic acquisition of high-resolution face images. Given the ability to acquire information at multiple scales, the challenges of relating this information across scales and interpreting this information become significant. Multiscale techniques open up a whole new area of research, including camera control, processing video from moving cameras, resource allocation, and task-based camera management in addition to challenges in performance modeling and evaluation.

### THE CONTEXTUAL EVENT DETECTION CHALLENGE

While detecting and tracking objects is a critical capability for smart surveillance, the most critical challenge in video-based surveillance (from the perspective of a human intelligence analyst) is interpreting the automatic analysis data to detect events of interest and identify trends. Current systems have just begun to look into automatic event detection. The area of context-based interpretation of the events in a monitored space is yet to be explored. Challenges here include: using knowledge of time and deployment conditions to improve video analysis, using geometric models of the environment and other object and activity models to interpret events, and using learning techniques to improve system performance and detect unusual events.

### THE LARGE SYSTEM DEPLOYMENT CHALLENGE

The basic techniques for interpreting video and extracting information from it have received a significant amount of attention. The next set of challenges deals with how to use these techniques to build large-scale deployable systems. Several challenges of deployment include minimizing the cost of wiring, meeting the need for low-power hardware for battery-operated camera installations, meeting the need for automatic calibration of cameras and automatic fault detection, and developing system management tools.

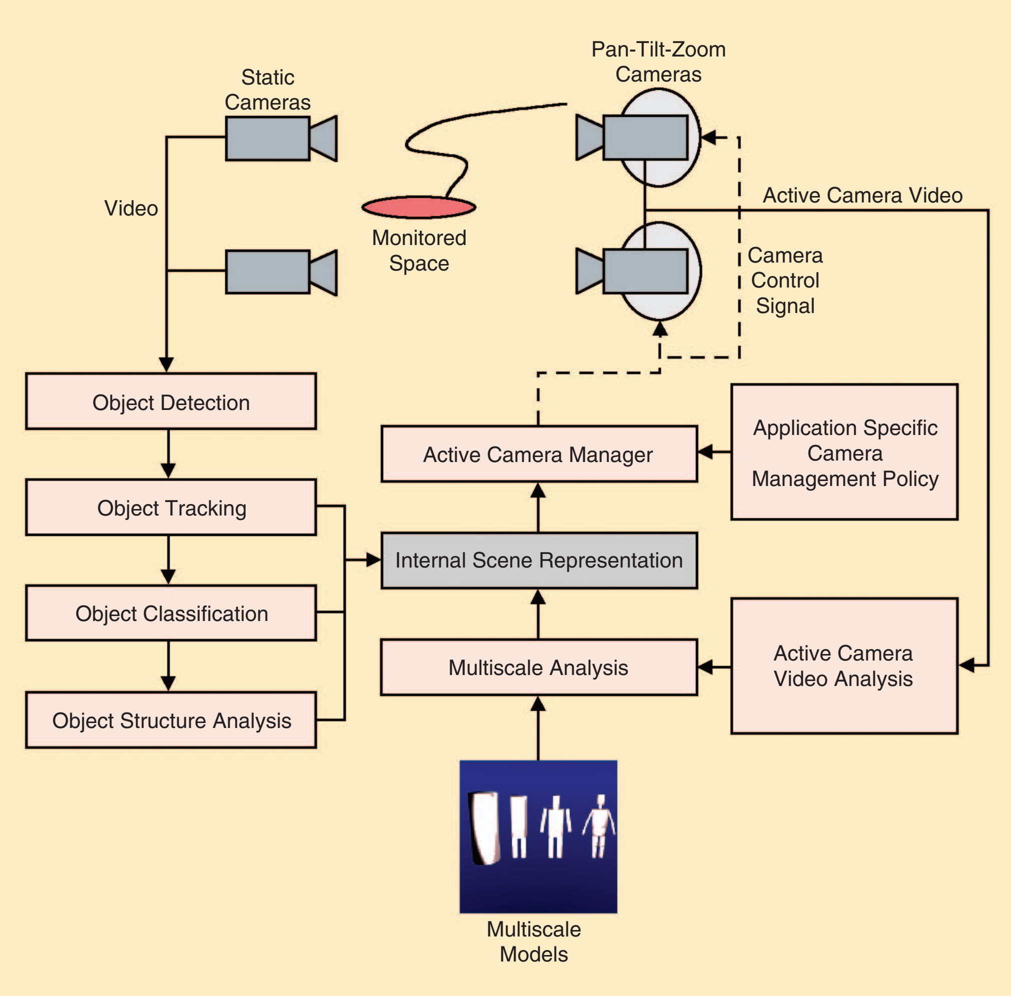
## COMPONENT TECHNOLOGIES FOR SMART SURVEILLANCE

Since the multiscale challenge incorporates the widest range of technical challenges, we present the generic architecture of a multiscale tracking system. The goal of a multiscale tracking system is to acquire information about objects in the monitored space at several scales in a unified framework. The architecture presented here provides a view of the interactions between the various components of such a system.

In Figure 1, the static cameras cover the complete scene of interest and provide a global view; the pan tilt zoom (PTZ) cameras are meant to obtain detailed or fine-scale information about objects of interest in the scene. The video from the static cameras is used to detect and track multiple objects in either two or three dimensions. Additionally, the fixed camera images can be used to extract additional information about the objects at a coarse level, like object class (person, car, truck) or object attributes (position of a person's head, velocity of the car, etc.). The coarse-scale information is used as a basis to "focus the attention of the PTZ cameras." The information from the PTZ cameras is then used to perform fine-scale analysis. For example, if the PTZ camera is directed towards a person, the fine-scale analysis could include face detection. The information from the coarse- and fine-scale analyses is combined in the internal scene representation. Specific instantiations of the multiscale architecture are presented in the following sections. We present the concepts that underlie several of the key techniques, including detection of moving objects in video, tracking in two and three dimensions, object classification, and object structure analysis. Our aim is to present the basic approach to each of these tasks. Research literature has many variants to the techniques presented.

### *OBJECT DETECTION*

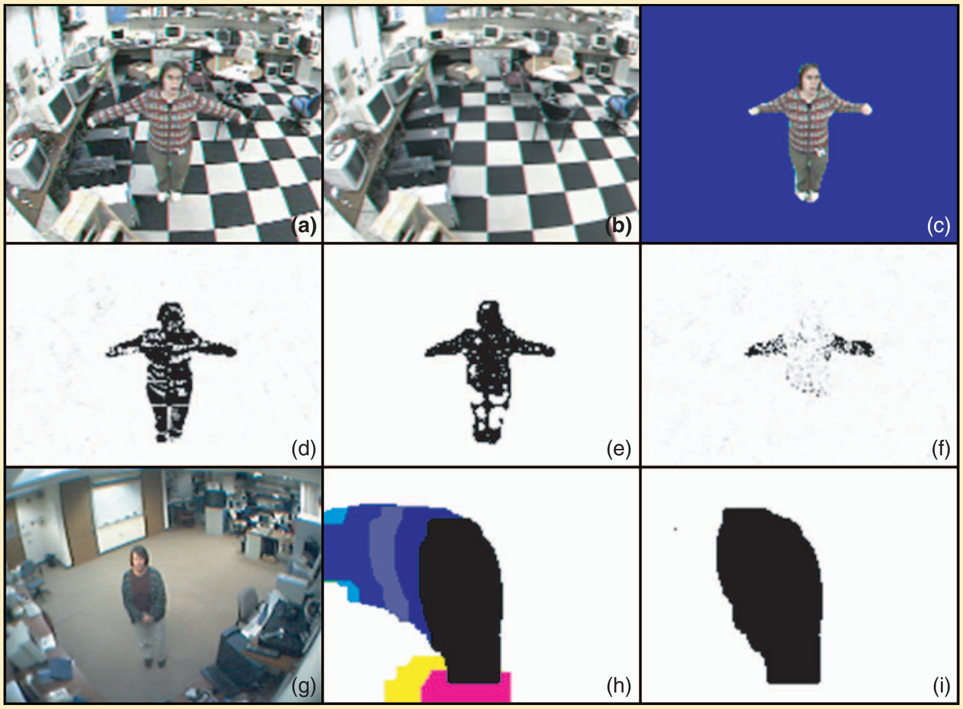
Object detection is the first stage in most tracking systems and serves as a means of focusing attention. There are two approaches to object detection. The first approach, called background subtraction, assumes a stationary background and treats all changes in the scene as objects of interest. The second approach, called salient motion detection, assumes that a scene will have many different types of motion, of which some types are of interest from a surveillance perspective. The following sections offer a short discussion of both approaches.



**[FIG1] Architecture of a generic multiscale tracking system. The system uses a combination of active cameras and multiscale models to address the issue of scale variations in the visual tracking applications. Different applications will implement a subset of this architecture.**

### ADAPTIVE BACKGROUND SUBTRACTION WITH HEALING

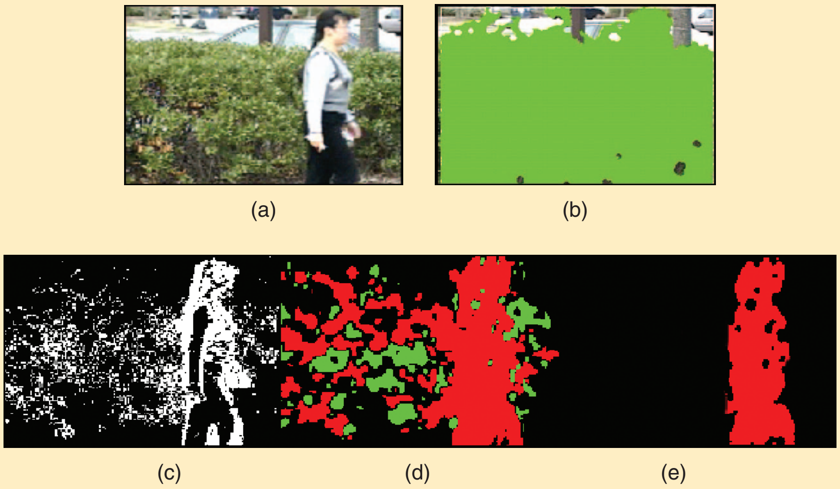
The background subtraction module combines evidence from differences in color, texture, and motion. Figure 2 shows the key stages in background subtraction. The use of multiple modalities improves the detection of objects in cluttered environments. The resulting saliency map is smoothed using morphological operators, and then small holes and blobs are eliminated to generate a clean foreground mask. The background subtraction module has a number of mechanisms to handle changing ambient conditions and scene composition. First, it continually updates its overall red/green/blue (RGB) channel noise parameters to compensate for changing light levels. Second, it estimates and corrects for automatic gain control (AGC) and automatic while balance (AWB) shifts induced by the camera. Third, it maintains a map of high-activity regions and slowly updates its background model only in areas deemed as relatively quiescent. Finally, it automatically eliminates occasional spurious foreground objects based on their motion patterns.



[FIG2] Background subtraction compares the current image (a) with a reference image (b) to find the changed regions (c) corresponding to objects of interest. Our system uses multiple modalities to locate change regions. It combines saliency from color shifts (d), differences in edge texture (e), and motion energy (f) to give a more robust segmentation. The system is able to incrementally build up and update a background model despite the presence of foreground objects. The person in the image (g) has just entered from the corner of the room. The system maintains a corresponding "quiescence" map (h) of where there have been no objects or motion for a while. From this, it generates a mask (i) of where the image is stable and hence where it is appropriate to adjust the background.

## SALIENT MOTION DETECTION

This is a complementary approach to background subtraction. Here, we approach the problem from a motion filtering perspective. Figure 3(a) shows a scene where a person is walking in front of a bush that is waving in the wind. Figure 3(b) shows the output of a traditional background subtraction algorithm which (per its design) correctly classifies the entire bush as a moving object. However, in this situation, we are interested in detecting the person as opposed to the moving bush. Our approach uses optical flow as the basis for detecting salient motion. We use a temporal window of $N$ frames (typically ten to 15) to assess the coherence of optic flow at each pixel over the entire temporal window. Pixels with coherent optical flow are labeled as candidates. The candidates from the motion filtering are then subjected to a region-growing process to obtain the final detection. These stages are shown in Figure 3(c)–(e).



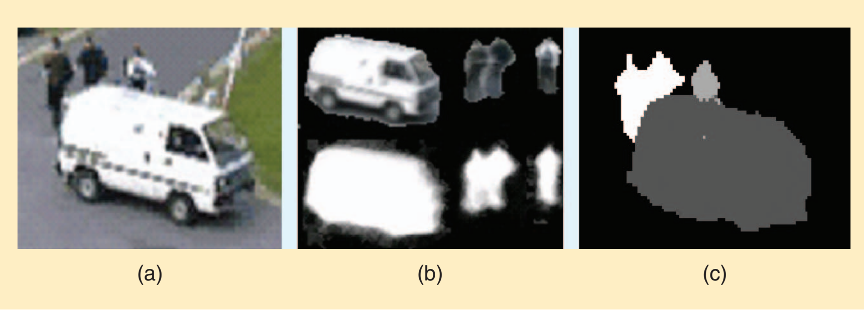
[FIG3] Illustration of salient motion detection. (a) Shows an image from a sequence with bushes in the background moving in the wind. (b) Shows the result obtained from our background subtraction algorithm where green regions indicate foreground demonstration of the limitation of background subtraction in regions of high environmental motion. (c)-(e) Show intermediate steps and results from the salient motion detection algorithm, (c) Shows a frame difference image. (d) Shows optical flow red patches show significant flow towards the right and green shows significant flow towards the left. (e) Shows the final result of segmentation where the bushes are classified as background and the person as foreground.

Background subtraction and salient motion detection are complementary approaches, each with its own strengths and weaknesses. Background subtraction is more suited for indoor environments where lighting is fairly stable and distracting motions are limited; salient motion detection is well suited to detect coherent motion in challenging environments with motion.

### *TWO-DIMENSIONAL OBJECT TRACKING*

Multi-object tracking aims to develop object trajectories over time by using a combination of the objects' appearance and movement characteristics. There are several challenges in multi-object tracking, with occlusion handling being one of the most critical. Our approach to multi-object blob tracking relies on appearance models that are image-based templates of object appearance. New appearance models are created when an object enters a scene. In every new frame, each of the existing tracks is used to try to explain the foreground pixels. The fitting mechanism used is correlation, implemented as minimization of the sum of absolute pixel differences between the detected foreground area and an existing appearance model. During occlusions, foreground pixels may represent the appearance of overlapping objects.

We use a maximum likelihood classification scheme to resolve foreground pixels into the component objects. The tracks are correlated in the order of their relative depth ordering (which has been computed in the previous frame). The correlation process is gated by the explanation map, which holds at each pixel the identities of the tracks explaining the pixels. Thus, foreground pixels that have already been explained by a track do not participate in the correlation process with models of the objects that are more distant. The explanation map is now used to update the appearance models of objects associated with each of the existing tracks. The explanation map is also used to determine the relative depth ordering. Regions of foreground pixels that are not explained by existing tracks are candidates for new tracks. A detailed discussion of the two-dimensional (2-D) multiblob tracking algorithm can be found in [18]. The 2-D multi-object tracker is capable of tracking multiple objects moving within the field of view of the camera, while maintaining an accurate model of the shape and color of the objects. Figure 4 illustrates an example of occlusion handling.



[FIG4] **Occlusion handling in object tracking: (a) original image with occlusion, (b) objects being tracked by the system, and (c) classification of pixels into models.**

### *THREE-DIMENSIONAL WIDE-BASELINE STEREO OBJECT TRACKING*

In several surveillance applications, it becomes necessary to determine the position of an object in the scene with reference to a three-dimensional (3-D) world coordinate system. This can be achieved by using two overlapping views of the scene and locating the same scene point in the two views. This approach to 3-D measurement is called stereo. There are two types of stereo: 1) narrow-baseline stereo, or stereo where the two cameras are placed close (a few inches) to each other, resulting in dense depth measurements at limited distance from the cameras, and 2) wide-baseline stereo where the two cameras are far apart (a few feet), resulting in a limited number of high-accuracy depth measurements where correspondences are available. In wide-area surveillance applications, wide-baseline stereo provides position information at large distances from the cameras, which is not possible with traditional stereo. Hence, we explore wide-baseline tracking. Figure 5 shows a block diagram of a 3-D tracker that uses wide-baseline stereo to derive the 3-D positions of objects. Video from each of the cameras is processed independently using the 2-D tracker described earlier, which detects objects and tracks them in the 2-D image.

The next step involves computing a correspondence between objects in the two cameras. The correspondence process is accomplished by using a combination of object appearance matching and camera geometry information. At every frame, we measure the color distance between all possible pairings of tracks from the two views. We use the Bhattacharya distance between the normalized color histograms of the tracks. For each pair, we also measure the triangulation error, which is defined as the shortest 3-D distance between the rays passing through the centroids of the appearance models in the two views. The triangulation error is generated using the camera calibration data. To establish correspondence, we minimize the color distance between the tracks from the view with the smaller number of tracks to the view with the larger number. This process can potentially cause multiple tracks from one view to be assigned to the same track in the other view. We use the triangulation error to eliminate such multiple assignments. The triangulation error for the final correspondence is thresholded to eliminate spurious matches that can occur when objects are just visible in one of the two views. Once a correspondence is available at a given frame, we need to establish a match between the existing set of 3-D tracks and 3-D objects present in the current frame. We use the component 2-D track identifiers of a 3-D track and match them against the component 2-D track identifiers of the current set of objects to establish the correspondence. The system also enables partial matches, thus ensuring a continuous 3-D track even when one of the 2-D tracks fails. Thus, the 3-D tracker is capable of generating 3-D position tracks of the centroid of each moving object in the scene. It also has access to the 2-D shape and color models from the two views that make up the track.

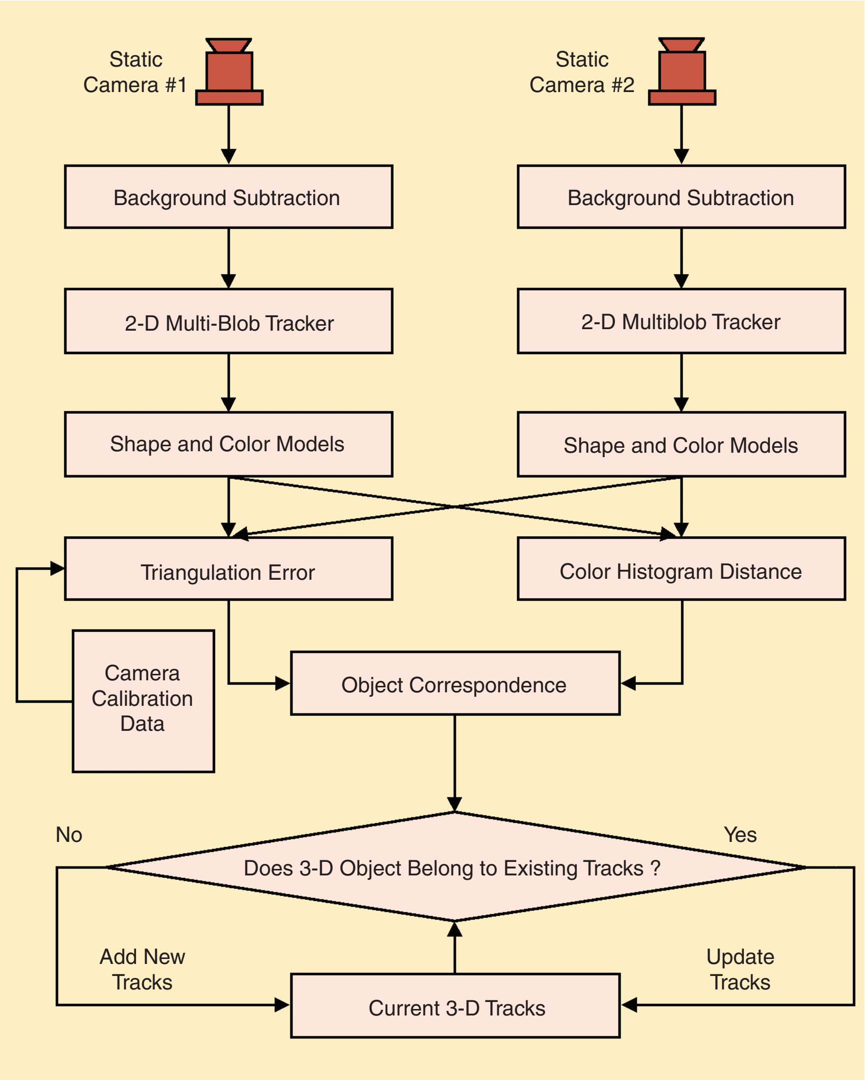
### *OBJECT CLASSIFICATION*

In several surveillance applications, determining the type of object is critical. For example, detecting an animal at a fence line bordering the woods may not be an alarm condition, whereas spotting a person there will definitely require an alarm. There are two approaches to object classification: an image-based approach and a video tracking-based approach. Presented below is a video tracking approach to object classification. This assumes that the objects of interest have been detected and tracked by an object tracking system. Image-based systems, such as face, pedestrian, or vehicle detection, find objects of a certain type without prior knowledge of the image location or scale. These systems tend to be slower than video tracking-based systems, which leverage current tracking information to locate and segment the object of interest.

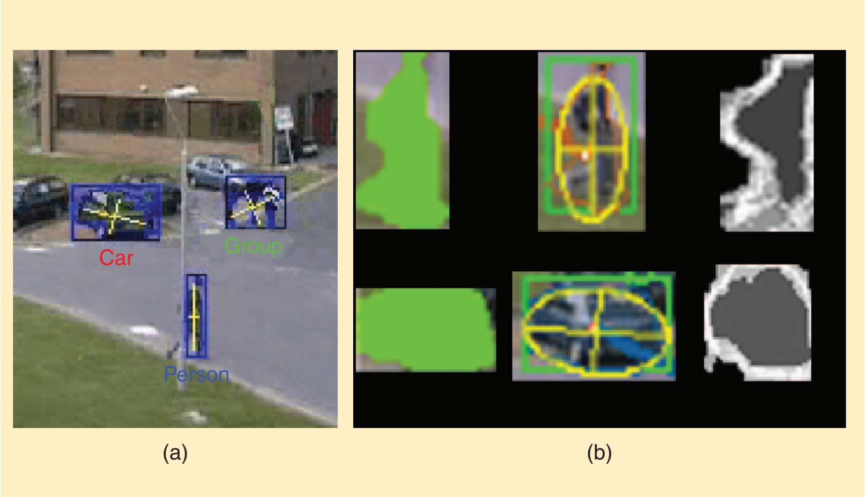
Video tracking-based systems use statistics about the appearance, shape, and motion of moving objects to quickly distinguish people, vehicles, carts, animals, doors opening/closing, trees moving in the breeze, etc. Our system (see Figure 6) classifies objects into vehicles, individuals, and groups of people based on shape features such as compactness, bounding ellipse parameters, and motion features (such as recurrent motion measurements, speed, and direction of motion). From a small set of training examples, we are able to classify objects in similar scenes using a Fisher linear discriminant to perform feature reduction, followed by a nearest neighbor classifier and temporal consistency information. Our classification system (not including tracking) runs at approximately 200 frames/s on a 2.4-GHz PC and accurately classified objects 90% of the time by track.

### *OBJECT STRUCTURE ANALYSIS: HEAD DETECTION*

Often, knowing that an object is present in the scene is not sufficient, and it

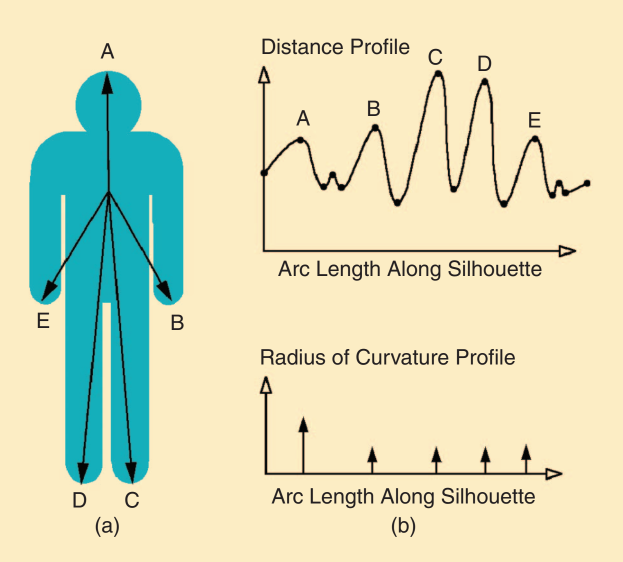


[FIG5] **Block diagram of a 3-D wide-baseline stereo tracker. This uses 2-D tracking in conjunction with camera calibration to track objects in 3-D.**



[FIG6] **(a) Output of object classification algorithm. (b) Intermediate steps in object classification.**

becomes necessary to analyze and locate parts of an object. For example, finding a person's head or the license plate of an automobile are important from an identification perspective. Our head detection technique uses the smoothed silhouette of the foreground object, as segmented using background subtraction. To interpret the silhouette, we use a simple human body model (see Figure 7) consisting of six body parts: head, abdomen, two hands, and two feet. First, we generate a one-dimensional (1-D) "distance profile," that is, the distance of each contour pixel from the contour centroid, following the contour clockwise. This distance profile is parsed into peaks and valleys based on the relative magnitudes of the successive extrema. The peaks of the distance transform are used to hypothesize candidate locations of the five body parts: the head, two feet, and two hands. Determination of the head among the candidate locations is currently based on a number of heuristics founded on the relative positions of the candidate locations and the curvatures of the contour at the candidate locations.


A
E
B
D
C
(a)
Distance Profile
A
B
C
D
E
Arc Length Along Silhouette
Radius of Curvature Profile
Arc Length Along Silhouette
(b)

**[FIG7] Various steps in object structure analysis: Head detection. (a) Shows the schematic of a human body with the five extreme points marked. The plots (b) and (c) show the distance to each of these extreme points from the centroid and the curvature at each of these points.**

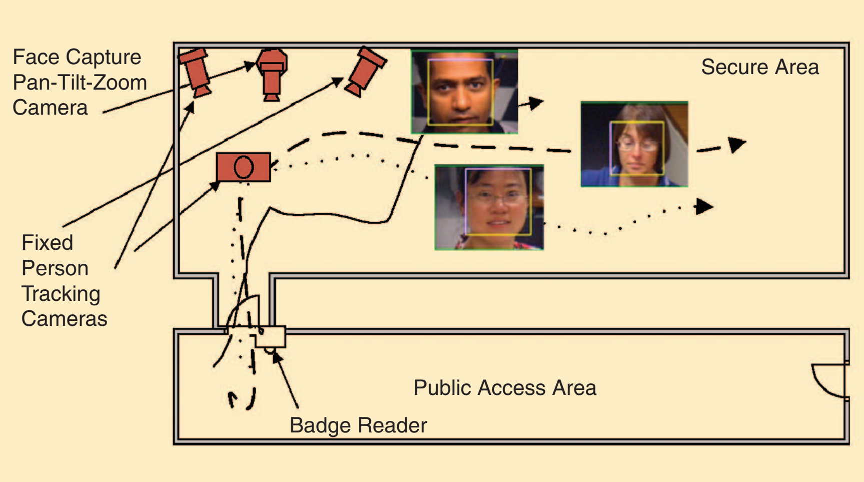
## FACE CATALOGER APPLICATION

The level of security at a facility is directly related to how well the facility can answer the question "who is where?" The "who" part of this question is typically addressed through the use of face images for recognition, either by an individual or a computer face recognition system. The "where" part of this question can be addressed through 3-D position tracking. The "who is where" problem is inherently multiscale; wide angle views are needed for location estimation and high-resolution face images are required for identification. An effective system to answer the question "who is where?" must acquire face images without constraining the users and must associate the face images with the 3-D path of the individual. The face cataloger uses computer-controlled PTZ cameras driven by a 3-D wide-baseline stereo tracking system. The PTZ cameras automatically acquire zoomed-in views of a person's head without constraining the subject to be at a particular location. The face cataloger has applications in a variety of scenarios where one would like to detect both the presence and identity of people in a certain space, such as loading docks, retail store warehouses, shopping areas, and airports.

Figure 8 shows the deployment of a face cataloger at the interface between a public access area and a secure area in a building. The secure area already has access control through a badge reader system; however, tailgating is a serious concern. A face cataloging system deployed in this area could help identify tailgaters and ensure that the identity of all people accessing the secure area is logged, irrespective of whether they badge-in.

The deployment consists of multiple fixed cameras that cover the area of the access point. These cameras are intended to detect and track the position of individual people in the scene. Additionally, there are PTZ cameras that are automatically controlled by the system to acquire high-resolution images of the faces of people passing through a space. The goal of the face cataloger is to acquire one face shot of each person who enters the space and associate the face image to the positional track of the person. Figure 8 schematically shows the association between the tracks of people and their pictures.

Figure 9 shows the schematic block diagram of the face cataloging system. Several components of this system have already been discussed. Here, we present how these component algorithms are integrated together. Given that the face



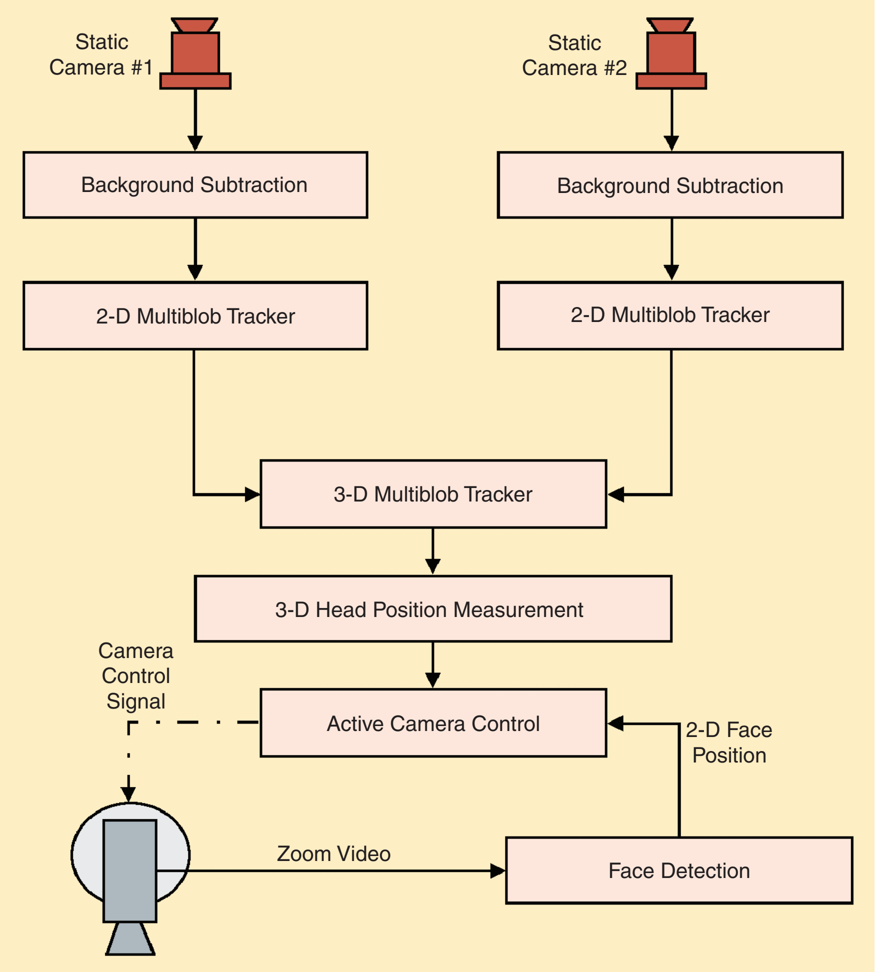
**[FIG8] A face cataloging system deployed at the interface between a secure area and public access area. The face catalog can be used for multiple purposes, including tailgating prevention.**

cataloger is a multicamera system, it inherently relies on the fact that all the cameras in the system are calibrated to a common world coordinate system. The calibration process is a one-time process performed at setup. Typically, it involves using a calibration pattern of known geometry or using physical measurements from the scene. The process uses images of the calibration pattern (in which feature points corresponding to known object geometry are manually selected) in conjunction with a few parameters supplied by the camera manufacturer [11], [17].

The following is a step-by-step description of the operation of the face cataloger system:

- *Step 1: 2-D Object Detection*—This step detects objects of interest as they move about the scene. The object detection process is independently applied to all the static cameras present in the scene.
- *Step 2: 2-D Object Tracking*—The objects detected in Step 1 are tracked within each camera field of view based on object appearance models.
- *Step 3: 3-D Object Tracking*—The 2-D object tracks are combined to locate and track objects in a 3-D world coordinate system. This step uses the 3-D wide-baseline stereo tracking discussed previously. The result of the 3-D tracking is an association between the same object as seen in two overlapping camera views of the scene.
- *Step 4: 3-D Head Detection*—To locate the position of the head in 3-D, we use the head detection technique described earlier. Given a 3-D track, the head is first detected in the corresponding 2-D views. The centroid of the head in the two views are used to triangulate the 3-D position of the head.
- *Step 5: Active Camera Assignment*—This step determines which of the available active cameras will be used for which task. Let us consider the example of a scene with three objects and a face cataloger system with two available active cameras. This step will employ an algorithm that uses an application-dependent policy to decide the camera assignment.
- *Step 6: 3-D Position-Based Camera Control*—Given the 3-D position of the head and a PTZ camera that has been assigned to the object, the system automatically steers the selected active camera to foveate in on the measured location of the head. There are several ways of controlling the pan-tilt and zoom parameters. For example, the zoom could be proportional to the distance of the object from the camera and inversely proportional to the speed at which the object is moving.
- *Step 7: Face Detection*—Once the 3-D position-based zoom has been triggered, the system starts applying face detection [12] to the images from the PTZ camera. As soon as a face is detected in the image, the control switches from 3-D position-based control to 2-D control based on face position.
- *Step 8: Face Detection-Based Camera Control*—Once the (frontal) face image is detected, the camera is centered on the face and the zoom is increased. The pan and tilt of the camera are controlled based on the relative displacement of the center of the face with respect to the center of the image. Given the intrinsic calibration parameters of the camera and the current zoom level (i.e., focal length), the relative image coordinate displacements are translated into desired (relative) pan/tilt angles. To avoid any potential instabilities in the feedback control strategy, we use a damping factor in the process.

Figure 10 shows images selected from the zoom sequence of the face cataloger. Figure 10(a)–(c) show the initial 3-D position-based control. Figure 10(d) shows the a box around the person's face, indicating that a face has been detected (Step 6). Figure 10(e)–(g) show the final stages of the zoom based on using the



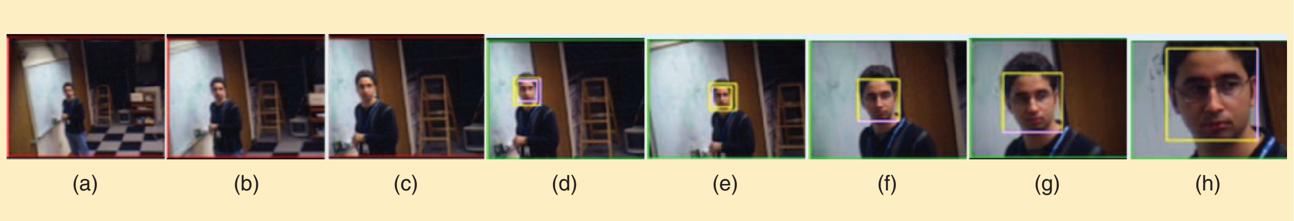
[FIG9] **Block diagram of the face cataloger.**

position and size of the face to control the PTZ of the camera. In a typical zoom sequence, the size of the face image will go from roughly ten pixels across to 145 pixels across the face with a resulting area zoom of 200. This clearly demonstrates the value of the multiscale approach. Details of the face cataloging technology can be found in [6].

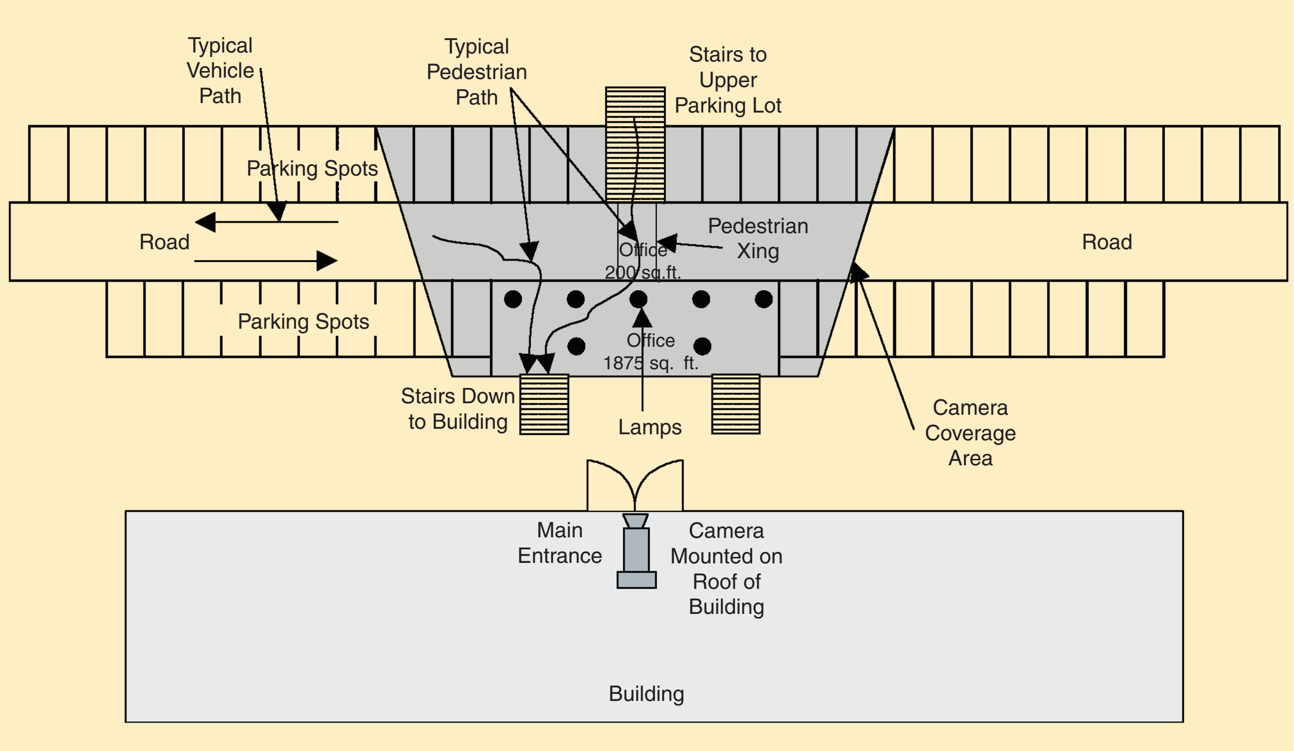
## LONG-TERM MONITORING AND MOVEMENT PATTERN ANALYSIS

Consider the challenge of monitoring the activity near the entrance of a building. Figure 11 shows the plan view of an IBM facility with a parking lot attached to the building. A security analyst would be interested in several types of activities, including finding cars speeding through the parking lot, finding cars that have been parked in loading zones, etc. Figure 12 shows the architecture of a system that can enable such queries through the use of smart surveillance technology.

The camera, which is mounted on the roof of the building, is wired to a central server. The video from the camera is analyzed by the smart surveillance engine to produce the viewable video index. The index is stored in the index database, which is a commercial database system, the IBM DB2. The video from the camera is also independently encoded by the video encoder and stored on a video server, the IBM Video Charger. An application can launch SQL queries against the index database to locate events of interest (such as all speeding cars between 9 a.m. and 10 a.m. on 13 January 2004). The events located in the index database are associated with a pointer to the actual video data on the video server, which can be used for browsing purposes.



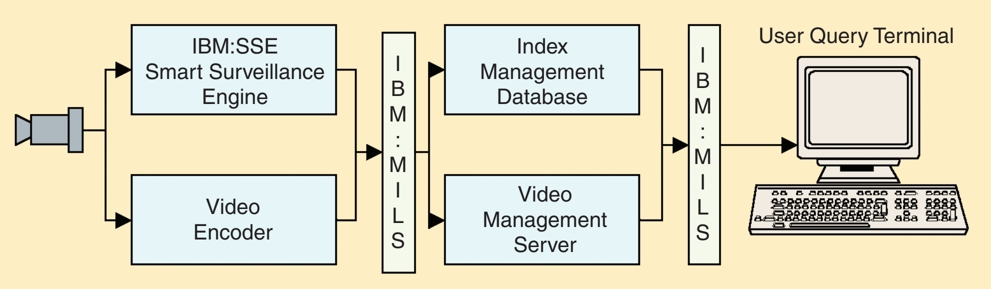
[FIG10] Face Cataloger zoom sequence. Images (a), (b), and (c) show successive stages as the camera zooms in on the person based on 3-D head position. Images (d), (e), (f), and (g) show the second stage of the zoom process, which uses face detection in a closed loop to acquire very high resolution images of the face. A video demonstration of this system can be viewed at http://www.research.ibm.com/peoplevision/facecateloger.html.



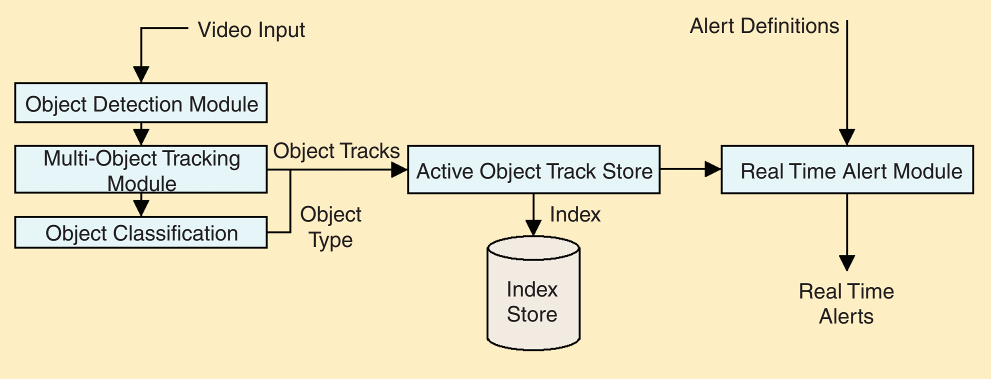
[FIG11] Plan view of the setup for monitoring the parking lot at an IBM facility. The camera, which is mounted on the building roof top, is monitoring the entrance area to the building and a segment of the parking lot attached to the building. Typical activity in the monitored area includes cars driving by, cars parking in a spot, people exiting/entering cars, and people walking in and out of the building.

The internal structure of the smart surveillance engine is shown in Figure 13. It uses some of the component technologies described earlier. The following analysis steps are performed on the video:

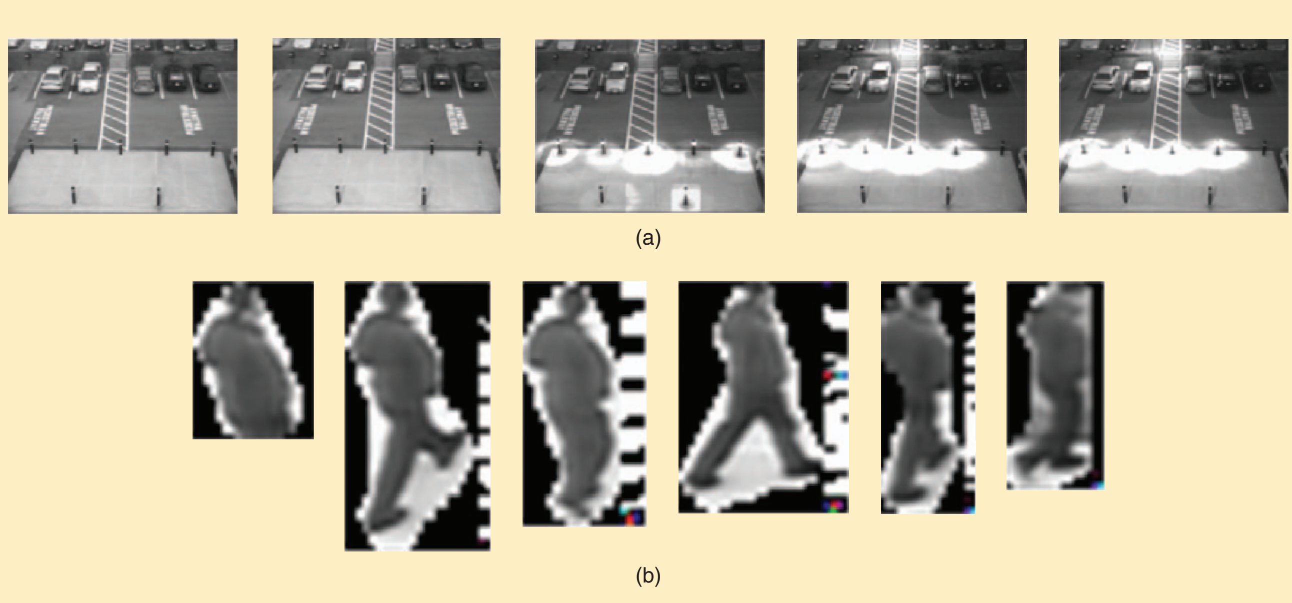
- *Step 1:* The first step is to detect objects of interest in the surveillance video. This step uses the object detection techniques described previously.
- *Step 2:* All the objects that are detected in Step 1 are tracked by the object tracking module, using the techniques described earlier. The key idea behind the tracking is to use the color information in conjunction with velocity-based prediction to maintain accurate color and shape models of the tracked objects.
- *Step 3:* Object classification is applied to all objects that are consistently tracked by the tracking module. The object classification currently generates three class labels, namely, vehicles, multiperson group, and single person.
- *Step 4:* Real-time alerts—these are based on criteria that set up by the user; examples include motion detection in specified areas, directional motion detection, abandoned object detection, object removal detection, and camera tampering detection.
- *Step 5:* Viewable video index (VVI)—this is a representation that includes time-stamped information about the trajectories of objects in the scene, other index data like size of the



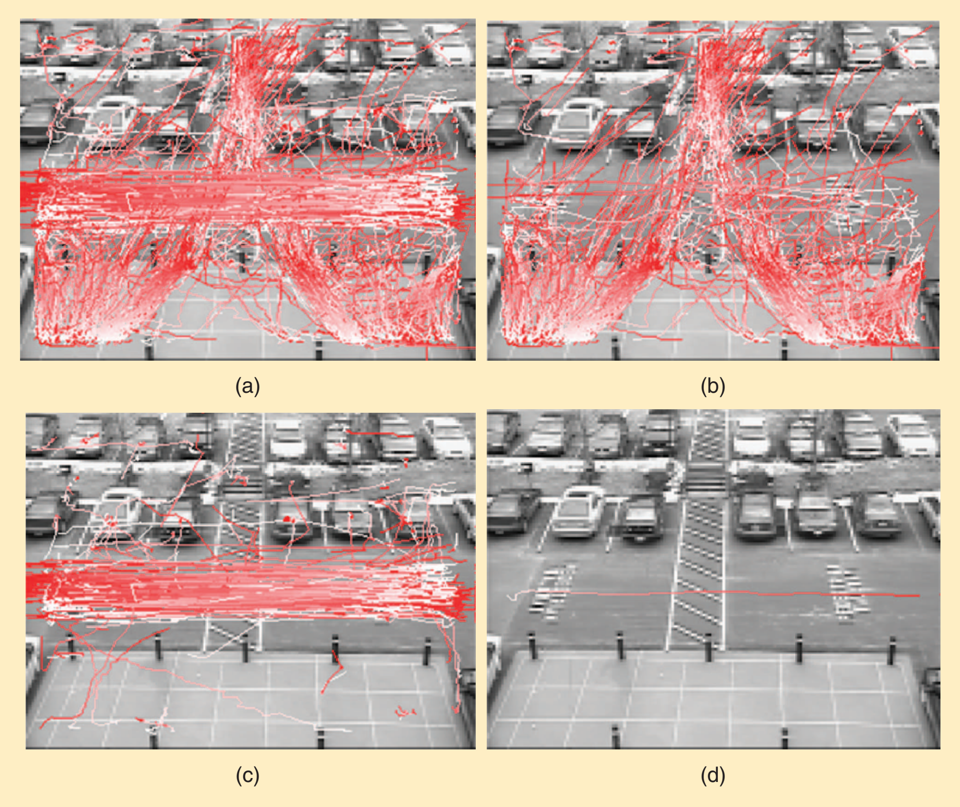
[FIG12] System architecture for long-term monitoring. The smart surveillance engine generates index data that is stored in a database. A movement pattern analysis program analyzes the index data in response to user queries.



[FIG13] Internal structure of the smart surveillance engine.



[FIG14] Visual proxies stored as part of the viewable video index. (a) Background models stored as the lighting in the scene changes going from day to night. (b) Object appearance model stored showing an objects appearance at various stages of its trajectory.



**[FIG15] (a) Tracks of all moving objects during the 24-hour monitoring period. (b) Tracks of objects with (50 pixels < Area < 1000 pixels) yielding only the people walking through the parking lot. (c) Tracks of objects with (1000 pixels < Area), yielding only vehicles. (d) Track of a speeding vehicle (Velocity > 10 pixels/frame ~ covering 10 m in 2 s ~ 11.25 mph). All tracks in the diagram are color coded, with tracks colored white at the beginning and gradually turning red as time progresses.**

object, type of object, etc. The viewable video index also incorporates appearance information in terms of the evolving background model and the appearance of objects through the lifetime. Figure 14 shows the visual portions of the index, which include background models that are indexed by time. A new background model is stored each time the background changes significantly [Figure 14(a)]. The index also stores foreground appearance models [Figure 14(b)]. These models are stored at intermittent points in the life of the foreground object in the scene. The index can be used to rerender the activities in the video.

### QUERYING THE VIEWABLE VIDEO INDEX

As the engine monitors the scene, it generates the VVI, which is stored into a relational database against which queries may be launched. There are several different types of basic queries that are supported by the database. Some of them are listed below:

- *Object Queries:* These are queries against the fixed properties of the object like object type and color. For example, show all cars that passed through this scene.
- *Temporal Queries:* These are queries against the time-varying properties of the object like motion properties, shape properties (for deformable objects), etc. For example, show all cars that drove through the scene in this direction.
- *Spatial Queries:* These are queries related to the positions of objects in the scene. For example, show all cars that drove through "this" part of the parking lot, where "this" is specified as an area in image coordinates.

The following is a presentation of results of queries performed against our experimental activity database. This test activity database was generated by monitoring activity from our experimental setup at an IBM facility as described in Figure 11. The database captured activity from 12 a.m. 13 January 2004 until 12 a.m. 14 January 2004. Figures 15 and 16 show the output of queries and the distribution of the arrival of people into the facility on 14 January. The index data can also be used to provide browse access to surveillance event data. Figure 17 shows a sample surveillance index browser.

### PERFORMANCE EVALUATION

Measuring the performance of smart surveillance systems is a very challenging task [5] due to the high degree of effort involved in gathering and annotating the ground truth as well as the challenges involved in defining metrics for performance measurement. Like any pattern recognition system, surveillance systems have two types of errors:

- *False Positives:* These are errors that occur when the system falsely detects or recognizes a pattern that does not exist in the scene. For example, a system that is monitoring a secure area may detect motion in the area when there is no physical object but rather a change in the lighting.
- *False Negatives:* These are errors that occur when the system does not detect or recognize a pattern that it is designed to detect. For example, a system monitoring a secure area may fail to detect a person wearing clothes similar to the scene background.

In this section, we present the various steps in evaluating an application like the face cataloging system. The ultimate goal of the face cataloger is to obtain good close-up head shots of people walking through the monitored space. The quality of the close-up face clips is a function of the accuracy of a number of underlying components. The following are potential sources of errors in the system:

- **Object Detection and Tracking Errors:**
  - *Frame-Level Object Detection Errors:* These are errors that occur when the detection system fail to detect or falsely detects an object.
  - *Scene-Level Object Detection Errors:* These are errors that occur when an object is  completely missed through its life in the scene or a completely nonexistent object is created by the system.
  - *2-D Track Breakage:* These errors occur when the tracker prematurely terminates a track and creates a new track for the same object.
  - *2-D Track Swap:* This error occurs when the objects being represented by a track get interchanged, typically after an occlusion.
  - *3-D Track Swap:* This error can occur due to errors in the inter-view correspondence process.
- **2-D Head Detection Errors:** These are errors in the position and size of the head detected in each of the 2-D views.
- **True Head Center Error:** Since we are detecting the head in two widely different views, the centers of the two head bounding boxes do not correspond to a single physical point and hence will lead to errors in the 3-D position.
- **3-D Head Position Errors:** These are errors in the 3-D position of the head due to inaccuracy in the camera calibration data.
- **Active Camera Control Errors:** These are errors that arise due to the active camera control policies. For example, the zoom factor of the camera is dependent on the velocity of the person; thus, any error in velocity estimation will lead to errors in the zoom control.
- **Active Camera Delays:** The delay in the control and physical motion of the camera will cause the close-up view of the head to be incorrect.

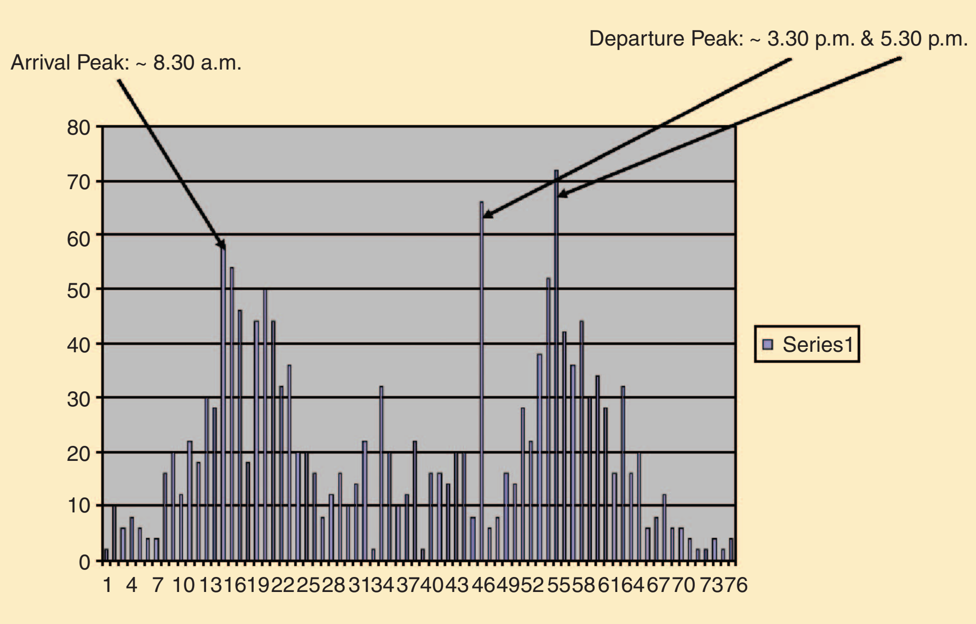
The above discussion illustrates how a variety of errors contribute to the final performance of the system. Below, we discuss the process we use to measure the performance of the most basic component of the system, namely object detection and tracking.

- *Test Data Collection and Characterization:* This involves collecting test sequences from one or more target environments. For example, the challenges in monitoring a waterfront at a port are very different from those of monitoring a crowded subway platform in New York City. Once data is collected from an environment, it is manually grouped into different categories based on environmental conditions (i.e., sunny day, windy day, etc.) for ease of interpreting the results of performance analysis.
- *Ground Truth Generation:* This is typically a very labor-intensive process and involves a human using a ground truth marking tool to manually identify various activities that occur in the scene. Our performance evaluation system uses a bounding box marked on each object every 30th frame while assigning a track identifier to each unique object in the scene.
- *Automatic Performance Evaluation:* The object detection and track systems are used to process the test data set and generate results. An automatic performance evaluation algorithm takes the test results and the ground truth data, compares them and generates both frame-level and scene-level object detection false positive and false negative results.

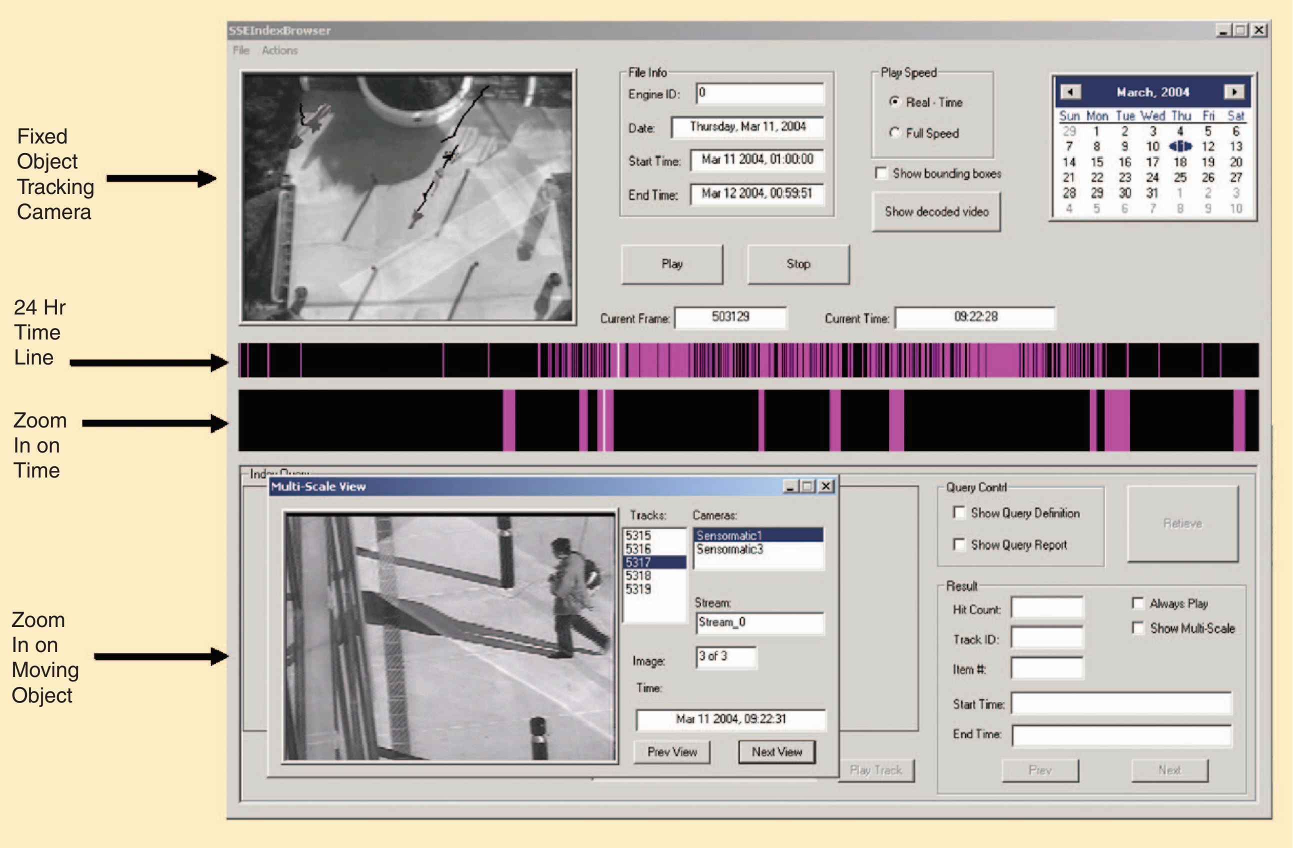
In summary, while a challenging task, performance evaluation of smart surveillance systems provides significant insights into the failure modes of the system. Our current efforts in performance evaluation show that our detection and tracking systems perform with roughly 70–95% accuracy. The chosen target environment for our system is an outdoor parking lot. Our efforts have also identified certain critical challenges, like camouflage and camera movement due to wind, which need to be addressed in order to improve performance.

## CONCLUSIONS AND FUTURE DIRECTIONS

Smart surveillance systems significantly contribute to situation awareness. Such systems transform video surveillance from a data acquisition tool to information and intelligence acquisition



**[FIG16] The temporal distribution of tracks corresponding to people (arriving and leaving the building). The tracks were selected using an area-based query [results presented in Figure 14(c)]. The peak of the arrivals is around 8:30 a.m., and the two significant departure peaks are around 3:30 p.m. and 5:30 p.m.**



File Info
Engine ID: 0
Date: Thursday, Mar 11, 2004
Start Time: Mar 11 2004, 01:00:00
End Time: Mar 12 2004, 00:59:51

Play Speed
● Real - Time
○ Full Speed
□ Show bounding boxes

March, 2004

| Sun | Mon | Tue | Wed | Thu | Fri | Sat |
|---|---|---|---|---|---|---|
| 29 | 1 | 2 | 3 | 4 | 5 | 6 |
| 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 21 | 22 | 23 | 24 | 25 | 26 | 27 |
| 28 | 29 | 30 | 31 | 1 | 2 | 3 |
| 4 | 5 | 6 | 7 | 8 | 9 | 10 |

Current Frame: 503129
Current Time: 09:22:28

Tracks: 5315<br>5316<br>5317<br>5318<br>5319
Cameras: Sensormatic1<br>Sensormatic3
Stream: Stream_0
Image: 3 of 3
Time: Mar 11 2004, 09:22:31

Query Contrl
□ Show Query Definition
□ Show Query Report

Result
Hit Count:
Track ID:
Item #:
Start Time:
End Time:
□ Always Play
□ Show Multi-Scale

[FIG17] Screen capture of a surveillance index browser showing the object tracking window, multiple time lines, and associated zoom pictures. The user can click on a particular event and browse the index from that point in time.

systems. Real-time video analysis provides smart surveillance systems with the ability to react to an activity in real-time, thus acquiring relevant information at much higher resolution. The long-term operation of such systems provides the ability to analyze information in a spatiotemporal context. As such systems evolve, they will be integrated both with inputs from other types of sensing devices and also with information about the space in which the system is operating, thus providing a very rich mechanism for maintaining situation awareness. Further details on all of the technologies described in this article can be found on our project Web site [14].

## AUTHORS

*Arun Hampapur* manages the Exploratory Computer Vision Group at the IBM T.J. Watson Research Center. He obtained his Ph.D. from the University of Michigan in 1995 in video data management systems. Before moving to IBM in 1997, he was leading the video effort at Virage Inc. At IBM Research, he leads an Adventurous Research project called PeopleVision. He has published over 30 papers on media indexing, video analysis, and video surveillance and holds seven U.S. patents. He serves on the program committees of several IEEE, ACM conferences and NSF review panels. He is a Senior Member of the IEEE.

*Lisa Brown* received her Ph.D. in computer science from Columbia University in 1995. For the past nine years, she has been a research staff member at the IBM T.J. Watson Research Center, where she is currently in the Exploratory Computer Vision Group. She has been an invited speaker and panelist at various workshops and has filed several patents. Her primary interests are in head tracking, head pose estimation, and, more recently, in object classification and performance evaluation.

*Jonathan Connell* is a research staff member at the IBM T.J. Watson Research Center in Hawthorne, New York. He graduated from MIT in 1989 with a Ph.D. in artificial intelligence doing work on behavior-based robotics. He was a past associate editor for *IEEE Transactions on Pattern Analysis and Machine Intelligence* and has taught in the Cognitive Science Program at Vassar College. He has 11 issued patents and has written books on robots, robot learning, and biometrics. His work has focused on computer vision for video search, biometric identification, and advanced user interfaces, and he is also interested in end-to-end AI systems and continues to pursue robotics in his spare time.

*Ahmet Ekin* received his Ph.D. degree in electrical and computer engineering from the University of Rochester, New York, in 2003. Since November 2003, he has been working as a senior research scientist in the Video Processing and Visual Perception group at Philips Research, Eindhoven, The Netherlands. Prior to joining Philips, he spent June-November 2003 at the IBM T.J. Watson Research Center, Hawthorne, New York, as a research co-op and helped design the face cataloger of IBM PeopleVision. He was also a consultant for Eastman Kodak Co., Rochester, New York. He was a member of the technical committee of IEEE ICIP 2001. He is a member of IEEE Signal Processing and Computer Societies.

***Norman Haas*** received the M.S. degree in computer science from Stanford University in 1978. He has been working at the IBM T.J. Watson Research Center since 1984, focusing on the areas of robotics, computer vision, and computational linguistics. He is currently a senior software engineer.

***Max Lu*** received his B.E. degree from Huazhong University of Science and Technology, Wuhan, China and his M.S. degree from National Laboratory of Pattern Recognition (NLPR), Chinese Academy of Sciences, Beijing, China. For many years, he has worked on building complex systems, such as machine tool computer numeric control (CNC) systems, image-based 3-D model capture systems, wafer defect detecting systems, etc. He is a contract software engineer of Exploratory Computer Vision Group, IBM Thomas J. Watson Research Center.

***Hans Merkl*** received his M.S. in mechanical engineering from Berufsakademie Stuttgart, Germany, in 1989. Prior to joining the IBM team, he was director of Decoder and Systems Products at ObjectVideo in Reston, Virginia, where he managed the development of ObjectVideo's MPEG-4 decoder efforts. Before joining ObjectVideo, he provided services as an independent consultant in Germany for various clients in the telecommunications and mobile industries. His clients included companies like Siemens AG, PICA, and Buehler.

***Sharath Pankanti*** obtained his Ph.D. from the Department of Computer Science, Michigan State University, in 1995. He joined the IBM T.J. Watson Research Center in 1995 as a postdoctoral fellow and became a research staff member in 1996. He worked on the IBM Advanced Identification Project until 1999. He is currently working on large-scale biometric indexing systems. He has coedited a book on biometrics, *Biometrics: Personal Identification* (Kluwer, 1999), and coauthored *A Guide to Biometrics* (Springer, 2004).

***Andrew Senior*** received his Ph.D. from Cambridge University in 1994. Since joining the IBM T.J. Watson Research Center in 1994, he has conducted research into handwriting, face, fingerprint, and audio-visual speech recognition. More recently, he has worked on tracking people with computer vision in the IBM PeopleVision project. He holds patents and has authored more than 40 scientific papers in these areas, including the recent book titled *Guide to Biometrics*.

***Chiao-Fe Shu*** received his Ph.D. from the Computer Science and Engineering Department of the University of Michigan in 1993. He is an expert architect, programmer, and researcher with over nine years of industrial experience. He cofounded Virage Inc. in 1994. His research covers the areas of oriented texture pattern analysis, classification, and segmentation, an in-situ wafer inspection system based on Fourier imaging, and multimedia indexing and retrieval. He has published extensively in his research areas and holds seven U.S. patents.

***Ying Li Tian*** received her Ph.D. from the Chinese University of Hong Kong in 1996 and her B.S. and M.S. degrees from TianJin University, China, in 1987 and 1990, respectively. After working at the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China, she joined the Robotics Institute at Carnegie Mellon University as a postdoctoral fellow. Currently, she is working at the IBM T. J. Watson Research Center, focusing on  the PeopleVison project. She has published more than 50 papers in journals and conferences. She is a Senior Member of IEEE.

## REFERENCES

[1] Blanz and Vetter, "Face recognition based on fitting 3D morphable model," *IEEE PAMI*, vol. 25, no. 9, pp. 1063–1074, Sept. 2003.

[2] R. Collins et al. "A system for video surveillance and monitoring," VSAM Final Report, Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-RI-TR-00-12, May 2000.

[3] Combat Zones That See, U.S. Government DARPA Project.

[4] M.W. Green, "The appropriate and effective use of security technologies in U.S. schools, A guide for schools and law enforcement agencies," Sandia National Laboratories, Albuquerque, NM, NCJ 178265, Sep. 1999.

[5] M. Greiffenhagen, D. Comaniciu, H. Niemann, and V. Ramesh, "Design, analysis and engineering of video monitoring systems: An approach and case study," *Proc. IEEE*, vol. 89, no. 10, pp. 1498–1517.

[6] A. Hampapur, S. Pankanti, A.W. Senior, Y-L. Tian, L. Brown, and R. Bolle, "Face cataloger: Multi-scale imaging for relating identity to location," in *Proc. IEEE Conf. Advanced Video and Signal Based Surveillance*, Miami, FL, 21–22 July 2003, pp. 13–20.

[7] A. Hampapur, L. Brown, J. Connell, M. Lu, H. Merkl, S. Pankanti, A. Senior, C. Shu, and Y. Tian, "The IBM smart surveillance system, demonstration," *Proc. IEEE,* CVPR 2004.

[8] Haritaoglu, "Harwood and Davis, W4: Real time surveillance of people and their activities," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 8, pp. 809–830, Aug. 2000.

[9] T. Horprasert, D. Harwood, and L. Davis, "A statistical approach for real-time robust background subtraction and shadow detection," in *Proc. IEEE Frame-Rate Workshop*, Kerkyra, Greece, 1999.

[10] Human ID at a Distance, U.S Government, DARPA Project.

[11] Intel. *Open Source Computer Vision Library (OpenCV)* [Online]. Available: http://www.intel. com/research/mrl/research/opencv

[12] R. Lienhart, A. Kuranov, and V. Pisarevsky, "Empirical analysis of detection cascades of boosted classifiers for rapid object detection," in *DAGM'03 25th Pattern Recognition Symposium*, Madgeburg, Germany, Sept. 2003, pp. 297–303.

[13] J. Phillips, P. Grother, R. Micheals, D.M. Blackburn, E. Tabassi, and M. Bone, "Face recognition vendor test 2002 P," in *Proc. IEEE Int. Workshop Analysis and Modeling of Faces and Gestures* (AMFG'03).

[14] IBM Research. *PeopleVision Project Home Page* [Online]. Available: http://www.research. ibm.com/peoplevision/

[15] C.S. Regazzoni, R. Visvanathan, and G.L. Foresti, "Scanning the issue/technology- special issue on video processing, understanding and communications in third generation surveillance systems," *Proc. IEEE*, vol. 89, no. 10, pp. 1419–1440, Oct. 2001.

[16] Remagnino, Jones, Paragios, and Regazzoni, *Video Based Surveillance Systems Computer Vision and Distributed Processing*. Norwell, MA: Kluwer , 2002.

[17] R.Y. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses," *IEEE J. Robot. Automat.*, vol. RA-3, no. 4, pp. 323–344, Aug. 1987.

[18] A. Senior, A. Hampapur, Y.-L. Tian, L. Brown, S. Pankanti, and R. Bolle, "Appearance models for occlusion handling," in *Proc. 2nd Int. Workshop Performance Evaluation of Tracking and Surveillance Systems in Conjunction with CVPR'01*, Dec. 2001.

[19] G. Stauffer, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 8, pp. 747–757, Aug. 2000.

[20] VACE: Video Analysis and Content Exploitation [Online]. Available: http://www.ic-arda.org/InfoExploit/vace/ SP